Examining reliability of seasonal to decadal sea surface temperature forecasts: the role of ensemble dispersion

Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion

Chun Kit Ho,[1] Ed Hawkins,[1] Len Shaffrey,[1] Jochen Bröcker,[2] Leon Hermanson,[3] James M. Murphy,[3] Doug M. Smith,[3] and Rosie Eade[3]

Received 15 August 2013; revised 9 October 2013; accepted 10 October 2013; published 11 November 2013.

[1] Useful probabilistic climate forecasts on decadal timescales should be reliable (i.e., forecast probabilities match the observed relative frequencies) but this is seldom examined. This paper assesses a necessary condition for reliability, which the ratio of ensemble spread to forecast error being close to one, for seasonal to decadal sea surface temperature retrospective forecasts from the Met Office Decadal Prediction System. Factors which may affect reliability are diagnosed by comparing this spread-error ratio for an initial condition ensemble and two perturbed physics ensembles for initialized and uninitialized predictions. At lead times less than 2 years, the initialized ensembles tend to be underdispersed and produce overconfident and hence unreliable forecasts. For longer lead times, all three ensembles are predominantly overdispersed. Such overdispersion is primarily related to excessive interannual variability in the climate model. These findings highlight the need to carefully evaluate simulated variability in seasonal and decadal prediction systems. **Citation:** Ho, C. K., E. Hawkins, L. Shaffrey, J. Bröcker, L. Hermanson, J. M. Murphy, D. M. Smith, and R. Eade (2013), Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion, *Geophys. Res. Lett.*, *40*, 5770–5775, doi:10.1002/2013GL057630.

1. Introduction

[2] Since skillful decadal climate forecasts could bring benefits to climate change adaptation planning, there has been significant development of such predictions in recent years, using global climate models (GCMs) initialized with atmospheric and oceanic observations [e.g., D. M. Smith et al., unpublished data, 2013]. Such decadal predictions are subject to uncertainties from different sources, such as the uncertainty in the initial state, the imperfect representation of the climate system by GCMs, and future changes in radiative forcing agents. Ensemble prediction systems have been developed to quantify some of these uncertainties by, for example, perturbing the initial conditions or model parameters of a single GCM [e.g., *Smith et al.*, 2010] or by combining different GCMs [e.g., *van Oldenborgh et al.*, 2012]. This raises the question of whether such systems can produce reliable probabilistic decadal climate predictions.

[3] Previous assessments of the quality of forecasts from ensemble decadal prediction systems have almost always focused on the accuracy of ensemble mean forecasts [e.g., *van Oldenborgh et al.*, 2012; *Ho et al.*, 2013]. However, a useful ensemble prediction system should also give *reliable* forecasts which means that the forecast probabilities match the observed relative frequencies. Evaluating the reliability of ensemble decadal predictions could aid forecast system development, for example, improving or informing ensemble generation. On seasonal timescales, several ensemble prediction systems tend to produce overconfident forecasts, and this has led to discussions about appropriate methods to increase the ensemble spread by sampling model uncertainty, initial condition uncertainty, and using stochastic physics [e.g., *Weisheimer et al.*, 2011; *Batté and Déqué*, 2012]. However, it is not yet clear whether similar conclusions will hold on decadal timescales. *Corti et al.* [2012] considered the reliability of ensemble decadal forecasts of multiyear land surface and sea surface temperatures (SSTs) on continental and ocean basin scales from a European Centre for Medium-Range Weather Forecasts (ECMWF) 54-member ensemble. Using reliability diagrams, they found that the ensemble was reliable overall, but that reliability was much reduced when the forced trends were removed.

[4] This paper evaluates the dispersion characteristics, a necessary condition for ensemble reliability, of SST forecasts from the UK Met Office Decadal Prediction System (DePreSys). In particular, we examine how the dispersion characteristics vary spatially and with forecast lead time from seasonal to decadal timescales. In addition, through a comparison of forecasts from three parallel DePreSys ensemble experiments, we aim to explore how model initialization, the use of perturbed physics, and the internal variability of the climate model contribute to the reliability of ensemble predictions.

2. Ensemble Experiments and Verifying Observations

[5] The Met Office Decadal Prediction System (DePreSys) [*Smith et al.*, 2010] is based on the third Hadley Centre coupled GCM (HadCM3) [*Gordon et al.*, 2000] which has a horizontal resolution of 2.5° × 3.75° in the

Additional supporting information may be found in the online version of this article.

[1]NCAS-Climate, Department of Meteorology, University of Reading, Reading, UK.
[2]Department of Mathematics and Statistics, University of Reading, Reading, UK.
[3]Met Office Hadley Centre, Exeter, UK.

Corresponding author: C. K. Ho, NCAS-Climate, Department of Meteorology, University of Reading, PO Box 243, Earley Gate, Reading, RG6 6BB, UK. (c.k.ho@reading.ac.uk)

0094-8276/13/10.1002/2013GL057630

atmosphere and 1.25° × 1.25° in the ocean. This paper considers three sets of retrospective forecast experiments, each consisting of nine ensemble members. Identical time-varying radiative forcings, derived from observed changes of greenhouse gases, aerosol, and solar irradiance, are used in each experiment. There are a total of 46 retrospective forecasts of global SSTs for each experiment, starting on 1 November of each year from 1960 to 2005, each extending to 9 years ahead [*Smith et al.*, 2010].

[6] 1. DePreSys ICE: An initial condition ensemble with the same physical parameters as the standard settings as HadCM3. For one of the nine members, atmospheric and oceanic analyses are assimilated as anomalies to create the initial conditions [*Smith et al.*, 2010]. The other eight members have different initial conditions which are created by adding small uncorrelated random SST perturbations.

[7] 2. DePreSys PPE: A perturbed physics ensemble consisting of different versions of HadCM3 with perturbations to poorly constrained physical parameters to sample this aspect of climate model uncertainty. One of the nine members uses the standard HadCM3 settings of physical parameters, while the other eight employ simultaneous perturbations of 29 atmospheric parameters [*Collins et al.*, 2011]. All nine members have the same initial conditions as in the first member of DePreSys ICE.

[8] 3. NoAssim PPE: A parallel ensemble to DePreSys PPE, but the initial conditions are taken from the appropriate points of transient simulations of the past climate, without assimilation of observations.

[9] The effect of model initialization on prediction skill and dispersion characteristics can be evaluated by comparing the DePreSys PPE and NoAssim PPE forecasts. The spread of DePreSys ICE is due to small differences in the initial conditions, and the additional effect of the perturbed parameters may be understood by comparing DePreSys PPE and ICE. Further details on the DePreSys experimental setup are given in the supporting information Text S1.

[10] In order to focus on the dispersion characteristics of forecasts of the internal variability, we remove the difference between observed and modeled long-term trend in SSTs by applying a linear bias adjustment, similar to that proposed by *Kharin et al.* [2012], to the DePreSys retrospective forecasts. This is performed on each grid box locally and for each lead time individually in a cross-validation manner. The details of this methodology are given in Text S2.

[11] HadISST global monthly interpolated SST data set [*Rayner et al.*, 2003] is used to verify the retrospective forecasts. These are interpolated onto the grid of HadCM3 using bilinear interpolation. The verification is only performed for grid boxes not covered by sea ice and from 35° S to 70° N due to the sparseness of observations over the southern oceans and near the Arctic.

3. Understanding Reliability Through Dispersion Characteristics

[12] A number of diagnostics can be used to assess the reliability of ensemble forecasts, such as reliability diagrams and rank histograms. However, their use may be limited by the small sample size available for verification, which is often the case for decadal forecast verification [*Corti et al.*, 2012]. It may also be impractical to study the spatial variation in reliability using these diagnostics as a large number of grid boxes are involved. Here we mainly consider a simple necessary condition for reliability based on the relationship between the intraensemble spread and the error of the ensemble mean forecast [*Weigel*, 2012]. This approach has been applied in assessing the need to calibrate ensemble predictions for weather [e.g., *Buizza*, 1997] and seasonal [e.g., *Weisheimer et al.*, 2011] timescales. For a reliable ensemble prediction system where the observation and the ensemble members are statistically indistinguishable, the average intraensemble variance $\sigma_e^2(\tau)$ and the mean squared error $\mathrm{MSE}(\tau)$ of the ensemble mean forecast for the same lead time τ should be related by

$$\sigma_e^2(\tau) = \frac{m}{m+1}\mathrm{MSE}(\tau) \tag{1}$$

where m is the number of ensemble members. We therefore consider the ratio of the time-averaged intraensemble standard deviation (σ_e) to the root-mean-squared error (RMSE) of the ensemble mean forecast, adjusted for the ensemble-size dependent factor in (1), for each grid box for different lead times. The ensemble is overdispersed (underdispersed) if this "spread-error ratio" ($\sqrt{10/9}\sigma_e$/RMSE) is greater (smaller) than one, and uncalibrated probabilistic forecasts produced from such an ensemble is expected to be unreliable. A bootstrapping approach similar to that employed in *Ho et al.* [2013] is used to estimate the sampling uncertainty of the spread-error ratio.

4. Results

4.1. Spread-Error Ratio for SSTs

[13] Figure 1 shows the spread-error ratio for the three ensembles for lead times of one season (the first winter—DJF), and 1, 3, and 9 years. Like many other seasonal forecast systems [e.g., *Weisheimer et al.*, 2011; *Batté and Déqué*, 2012], DePreSys ICE is underdispersed nearly everywhere for the first season (top row). This underdispersion, which often corresponds to overconfident and hence unreliable forecasts, is somewhat mitigated when considering DePreSys PPE, demonstrating the benefits of the perturbed physics approach to sample aspects of model uncertainty and potentially produce more reliable predictions. However, large regions of underdispersion remain, particularly in the tropical Pacific. Interestingly, NoAssim PPE is generally *overdispersed* for this season, suggesting that the initialization is the primary reason for underdispersion.

[14] Considering the first annual mean (second row), the picture changes. Although all the ensembles are underdispersed in the tropical Pacific, in the extratropics, they are overdispersed. By year 9 (bottom row), the patterns of the spread-error ratio converge across the ensembles, with 65 to 75% of grid points showing significant (at the 10% level) overdispersion, which corresponds to underconfident and hence also unreliable forecasts. The North Atlantic is particularly overdispersed, with the spread being up to a factor of 2 too large. The small number of grid boxes (1 to 2% of the total) with the ratio significantly less than one is confined to the tropical Pacific.

[15] Initially, this overdispersion may seem surprising, but this ensemble comparison indicates that it is not the initialization process itself, or the perturbed physics, which is responsible for the long lead time overdispersion in DePreSys PPE.

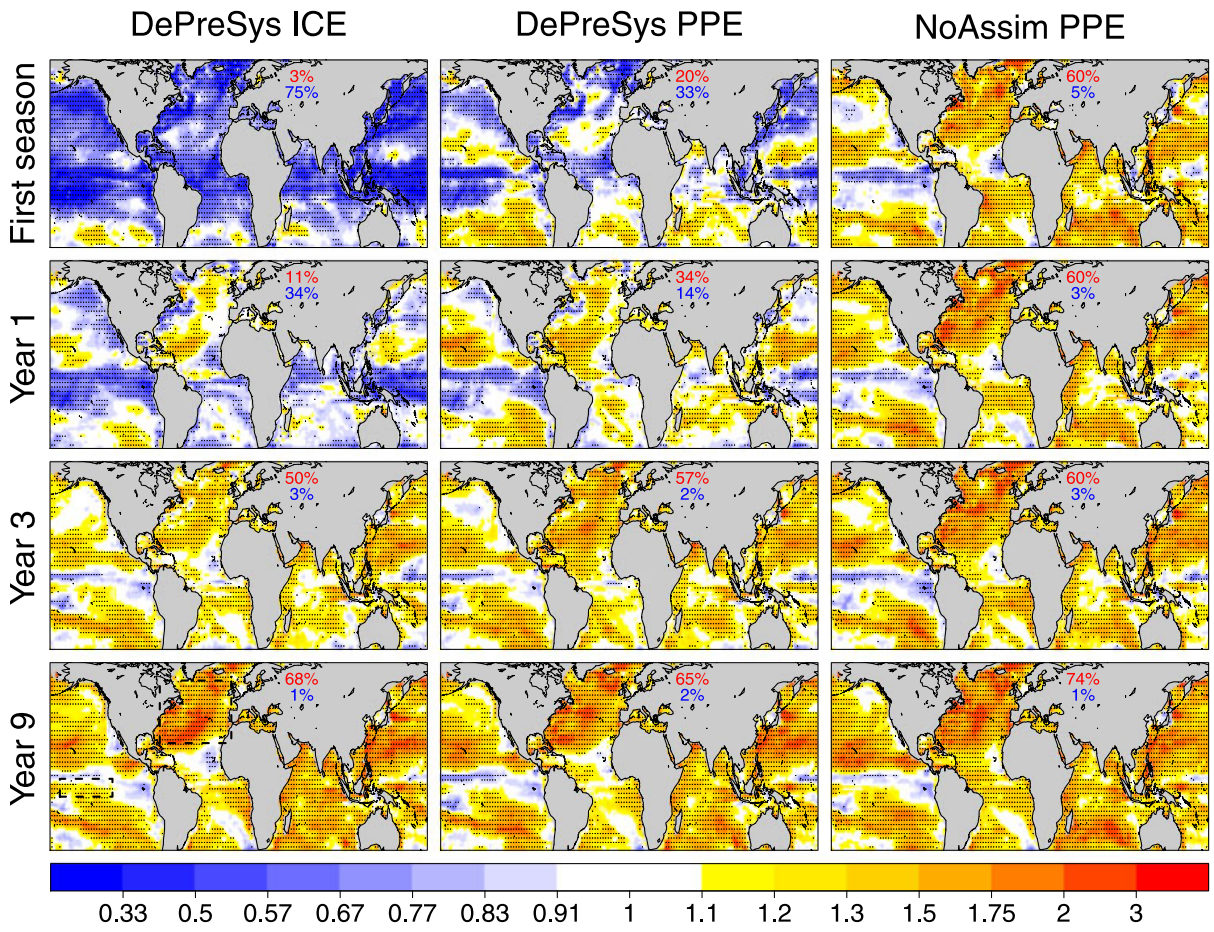

Figure 1. Spread-error ratio (ratio of mean intraensemble standard deviation and root-mean-squared error of ensemble mean forecasts, adjusted for ensemble size) for the three ensemble experiments for four different forecast lead times. Stippled areas indicate where the ratio is significantly different from one at the 10% level. The number in red (blue) in each panel is the proportion of grid boxes with the spread-error ratio significantly greater (smaller) than one. The boxes in the bottom left panel indicate the regions examined in Figures 3 and S2.

4.2. Diagnosing the Skill and Spread-Error Ratio

[16] The dispersion patterns can be partly understood by separating the spread-error ratio into its different components and comparing pairs of ensembles (Figure 2). In year 1, the DePreSys PPE ensemble has a larger spread than the ICE ensemble by 10 to 30% (left column of Figure 2a), but these differences reduce with lead time. Meanwhile, comparing DePreSys and NoAssim PPE (right column of Figure 2a) demonstrates that initialization significantly reduces forecast spread. However, this effect also decays over time, but more slowly in the extratropics. By year 9, there is very little difference between the ensemble spreads.

[17] It is also interesting to consider the differences in RMSE, a measure of forecast skill (Figure 2b). At lead times of 1 and 3 years, DePreSys PPE is more skillful (with smaller RMSE) overall than DePreSys ICE, especially in the Indian and Pacific Oceans. These differences remain for a few years, but by year 9, DePreSys ICE appears more skillful, especially in the Atlantic. Also, the benefit of initialization on skill is clear for year 1 (right column of Figure 2b), with around 50% less RMSE in many tropical regions. At year 3, such benefits remain for the North Atlantic only, but at year 9 the PPE initialization seems to produce less skillful forecasts than NoAssim PPE in most regions. In the North Atlantic Current region and parts of the western North Pacific, however, the RMSE for DePreSys PPE is larger than NoAssim PPE even at year 1. We note that if the more conventional "mean bias" correction is applied to the retrospective forecasts instead of trend adjustments, the difference in RMSE between DePreSys and NoAssim is somewhat smaller (Text S3 and Figure S1).

[18] We have so far verified forecasts for lead times of one season and three individual years, all with start dates from every year. In the decadal prediction literature, multiyear average predictions and forecasts with less frequent start dates are often considered [*Goddard et al.*, 2013]. In our case, the results for lead times of 2–5 years and 6–9 years (Figure S5) are similar to that for year 3 and year 9 in Figure 1. Also, a similar spatial pattern of the spread-error ratio is obtained when we perform the verification on a subset of forecasts with start dates every 5 years instead (Figure S6), indicating that this metric of reliability can be applied to simulations performed as per the Coupled Model Intercomparison Project Phase 5 (CMIP5) protocol.

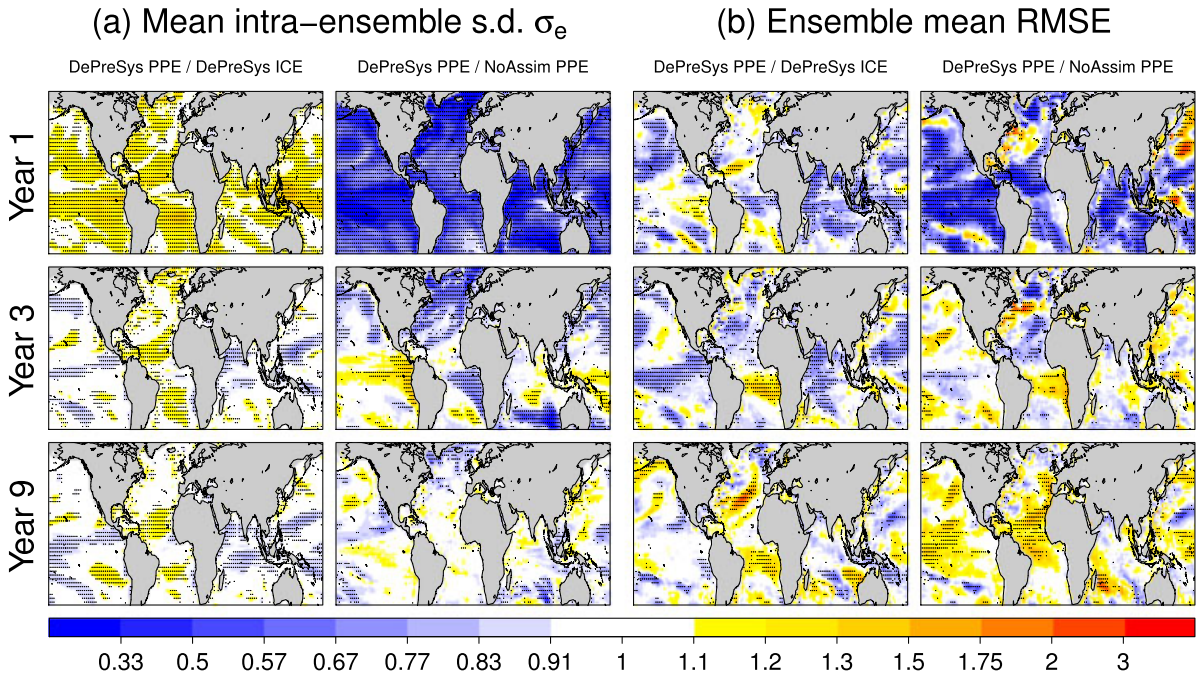

Figure 2. Comparison of (a) mean intraensemble standard deviation (σ_e) and (b) root-mean-squared error (RMSE) of ensemble mean forecasts for DePreSys PPE, DePreSys ICE, and NoAssim PPE, as ratios are indicated on the top of each column. In Figure 2a, blue shades mean that DePreSys PPE has a smaller spread. In Figure 2b, blue shades mean that the ensemble mean DePreSys PPE forecasts are more accurate. Stippled areas indicate where the ratio is significantly different from one at the 10% level. The mean intraensemble sd and RMSE for each ensemble are shown in Figures S3 and S4.

4.3. Regional Analysis—North Atlantic and Nino 3.4

[19] We now examine the dispersion characteristics of the three ensembles for retrospective forecasts of two specific area averages: the North Atlantic and the Nino 3.4 region. Figure 3 shows how the spread and RMSE vary as a function of lead time. For the North Atlantic (Figure 3a), NoAssim PPE is overdispersed for all lead times, consistent with the spatial maps shown in Figure 1. In contrast, the spread of the two initialized ensembles, DePreSys PPE and DePreSys ICE, increases gradually with lead time and remains smaller than that of NoAssim PPE up to year 9. The RMSE for DePreSys PPE and DePreSys ICE also increases with lead time, but more slowly than the spread, so the two ensembles become overdispersed.

[20] For the Nino 3.4 region (Figure 3b), there is also little variation in the spread of NoAssim PPE with lead time. The spread of DePreSys PPE is about 45% lower than that of NoAssim PPE in year 1, but they become comparable by year 3. The spread of DePreSys ICE is larger than that of DePreSys PPE and NoAssim PPE, which indicates the difference in the properties of simulated El Niño-Southern Oscillation among the perturbed physics variants [*Toniazzo*

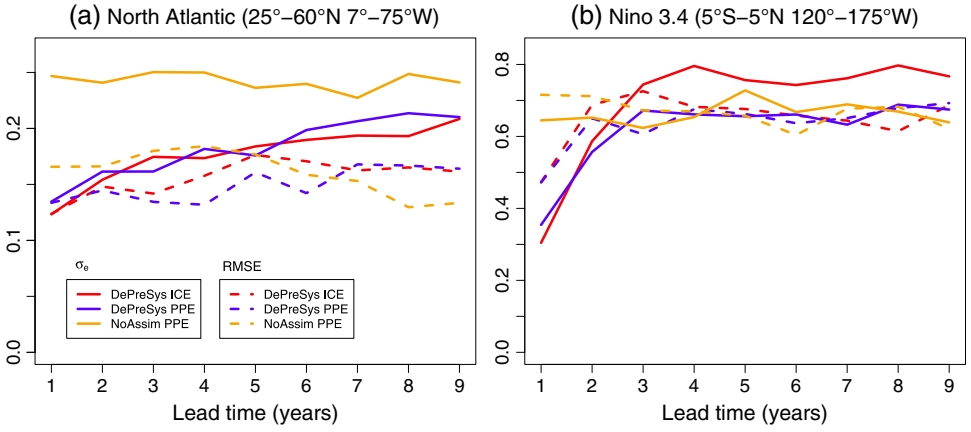

Figure 3. Mean intraensemble standard deviation (σ_e in K; solid line) of average SSTs and root-mean-squared error of ensemble mean (in K; dashed line) average SSTs in (a) North Atlantic region and (b) Nino 3.4 region as a function of forecast lead time for three sets of ensemble runs as indicated in the legend.

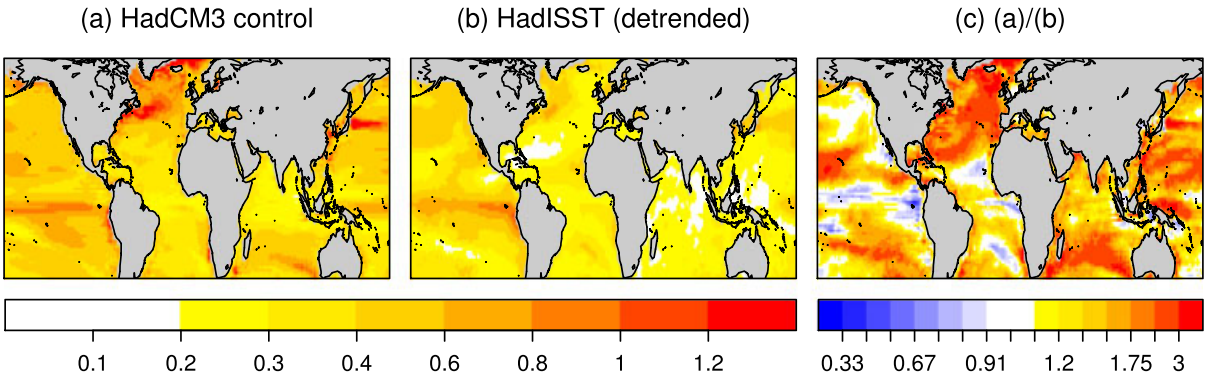

Figure 4. Standard deviation (in K) of (a) HadCM3 control integration and (b) linearly detrended HadISST during the verification period. (c) The ratio of Figure 4a to Figure 4b.

et al., 2008]. The RMSE of DePreSys PPE is also lower than that of NoAssim PPE at short lead times, but becomes comparable from year 4 onward. The impact of model initialization on both the spread and skill of the ensemble forecasts persists for a shorter time for Nino 3.4 compared to the North Atlantic region. All three ensembles are underdispersed for years 1 and 2, but DePreSys ICE becomes overdispersed at longer lead times, while the DePreSys PPE and NoAssim PPE have no clear signs of overdispersion or underdispersion.

[21] We also consider rank histograms [*Weigel*, 2012] for these regional average forecasts as an additional diagnostic for reliability (Text S4 and Figure S2). The results are noisy due to the small sample size, but they are generally consistent with that described above.

4.4. Why Do the Ensembles Become Overdispersed?

[22] Finally, we consider the reason for the overdispersion found in the ensembles: climatological variance. As noted by *Johnson and Bowler* [2009], for a reliable system it is also necessary to have the climatological variance of the observations and the underlying model to be the same, in addition to fulfilling the spread-error ratio condition (1). Figure 4 compares the standard deviation (sd) of the control integration of HadCM3, the climate model on which DePreSys is based, with the sd of linearly detrended HadISST during the verification period. The sd of the control run is larger than that of HadISST in most places, by a factor of 2 or more in parts of the North Atlantic and North Pacific (Figure 4c). This pattern is similar to the overdispersion seen in Figure 1 (bottom row), suggesting that the excessive variability in the climate model contributes to the general overdispersion for DePreSys ensembles in these regions. The tropical Pacific is the only region where the forecasts tend to be underdispersed at long lead times. In this region, the variability in the ensemble is more similar to the observations. However, note that our assessment has used a single observational data set (HadISST) which is subject to possible errors and uncertainties in its variability characteristics.

[23] As a further test, we have repeated the verification for DePreSys PPE using a perfect model approach where the transient simulations of each PPE member are used in turn as the verifying observations (Text S5). Overall the average spread-error ratios for the nine verifications at long lead times are close to one in most places. This confirms that the overdispersion is related to the differences in internal variability between model simulations and observations. However, there is a wide range of behaviors across the different ensemble members (Figure S7). Further work will attempt to determine whether any combination of parameter settings is producing excessive variability.

5. Conclusions

[24] This paper has assessed the dispersion characteristics of three ensemble decadal SST predictions from the Met Office Decadal Prediction System (DePreSys) in order to understand their capability to produce reliable probabilistic forecasts. The main findings are the following.

[25] 1. Dispersion characteristics of decadal prediction ensembles for SSTs vary considerably both spatially and with forecast lead time.

[26] 2. For lead times of less than 2 years, the initialized ensembles tend to be underdispersed and give overconfident and hence unreliable forecasts, especially in the tropics, consistent with many previous studies on this timescale.

[27] 3. For longer lead times, up to 9 years, the ensembles become overdispersed in most regions and thus give underconfident and also unreliable forecasts. Such overdispersion is related to excessive underlying variability in the climate model.

[28] These results have important implications. First, choices in the ensemble design for decadal predictions (e.g., stochastic or perturbed physics approaches) have been partly motivated by the underdispersion seen on seasonal timescales. However, our results indicate that the variability of the underlying climate model is at least as important as the ensemble perturbation scheme in producing reliable decadal climate forecasts. Evaluating the simulated variability during model development is therefore essential. Second, the excessive variability of SSTs in the climate model may affect the predictability over land on the decadal timescale.

[29] Our assessment has focused on the ratio of intra-ensemble spread and the error of the ensemble mean forecast. While this simple diagnostic should not be viewed as a complete evaluation of reliability, which would require a flow-dependent perspective, it is clearly helpful in identifying where and for what lead times the ensemble decadal forecasts are overdispersed or underdispersed and hence unreliable, even with a limited number of available verification cases.

[30] **Acknowledgments.** We thank Jon Robson, David Sexton, and Rowan Sutton for useful comments and suggestions. We also thank Antje Weisheimer, Susanna Corti, and the anonymous reviewers for their comments on this manuscript. We thank VALOR, a RAPID-WATCH project for providing one of the data sets used in this study. This work has received funding from the European Community's 7th Framework Programme (FP7) under grant agreements GA212643 (THOR) and GA303378 (SPECS).

[31] The Editor thanks two anonymous reviewers for their assistance in evaluating this paper.

References

Batté, L., and M. Déqué (2012), A stochastic method for improving seasonal predictions, *Geophys. Res. Lett.*, *39*, L09707, doi:10.1029/2012GL051406.

Buizza, R. (1997), Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system, *Mon. Weather Rev.*, *125*(1), 99–119.

Collins, M., B. B. B. Booth, B. Bhaskaran, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb (2011), Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles, *Clim. Dynam.*, *36*, 1737–1766, doi:10.1007/s00382-010-0808-0.

Corti, S., A. Weisheimer, T. N. Palmer, F. J. Doblas-Reyes, and L. Magnusson (2012), Reliability of decadal predictions, *Geophys. Res. Lett.*, *39*, L21712, doi:10.1029/2012GL053354.

Goddard, L., et al. (2013), A verification framework for interannual-to-decadal predictions experiments, *Clim. Dynam.*, *40*, 245–272, doi:10.1007/s00382-012-1481-2.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood (2000), The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments, *Clim. Dynam.*, *16*(2-3), 147–168, doi:10.1007/s003820050010.

Ho, C. K., E. Hawkins, L. Shaffrey, and F. M. Underwood (2013), Statistical decadal predictions for sea surface temperatures: A benchmark for dynamical GCM predictions, *Clim. Dynam.*, *41*, 917–935, doi:10.1007/s00382-012-1531-9.

Johnson, C., and N. Bowler (2009), On the reliability and calibration of ensemble forecasts, *Mon. Weather Rev.*, *137*, 1717–1720, doi:10.1175/2009MWR2715.1.

Kharin, V. V., G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W. S. Lee (2012), Statistical adjustment of decadal predictions in a changing climate, *Geophys. Res. Lett.*, *39*, L19705, doi:10.1029/2012GL052647.

Rayner, N., D. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *J. Geophys. Res.*, *108*, 4407, doi:10.1029/2002JD002670.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife (2010), Skilful multi-year predictions of Atlantic hurricane frequency, *Nat. Geosci.*, *3*, 846–849, doi:10.1038/NGEO1004.

Smith, D. M., et al. (2013), Real-time multi-model decadal climate predictions, *Clim. Dynam.*, doi:10.1007/s00382-012-1600-0.

Toniazzo, T., M. Collins, and J. Brown (2008), The variation of ENSO characteristics associated with atmospheric parameter perturbations in a coupled model, *Clim. Dynam.*, *30*, 643–656, doi:10.1007/s00382-007-0313-2.

van Oldenborgh, G., F. Doblas-Reyes, B. Wouters, and W. Hazeleger (2012), Decadal prediction skill in a multi-model ensemble, *Clim. Dynam.*, *38*, 1263–1280, doi:10.1007/s00382-012-1313-4.

Weigel, A. P. (2012), Ensemble forecasts, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed., chap. 8, edited by D. B. Stephenson and I. T. Jolliffe, pp. 141–166, Wiley-Blackwell, Oxford, U. K.

Weisheimer, A., T. N. Palmer, and F. J. Doblas-Reyes (2011), Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles, *Geophys. Res. Lett.*, *38*, L16703, doi:10.1029/2011GL048123.